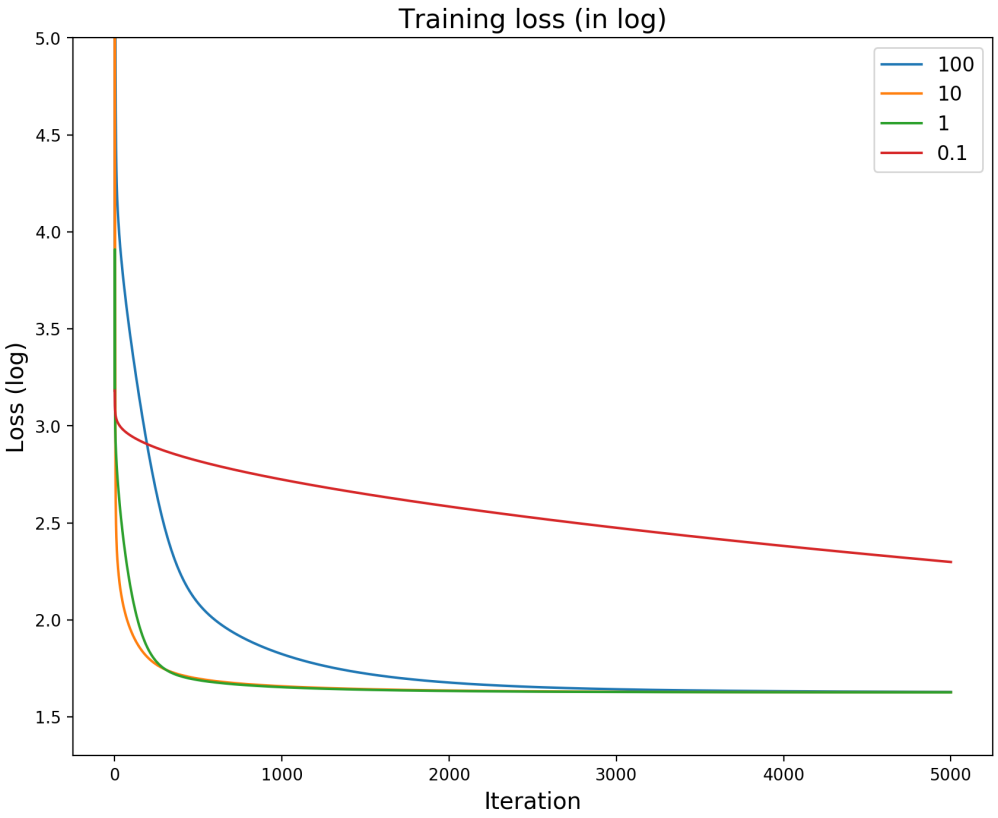
ML2020SPRING HW1 Report

學號：R08946015 系級：資料科學碩一 姓名：陳鈞廷

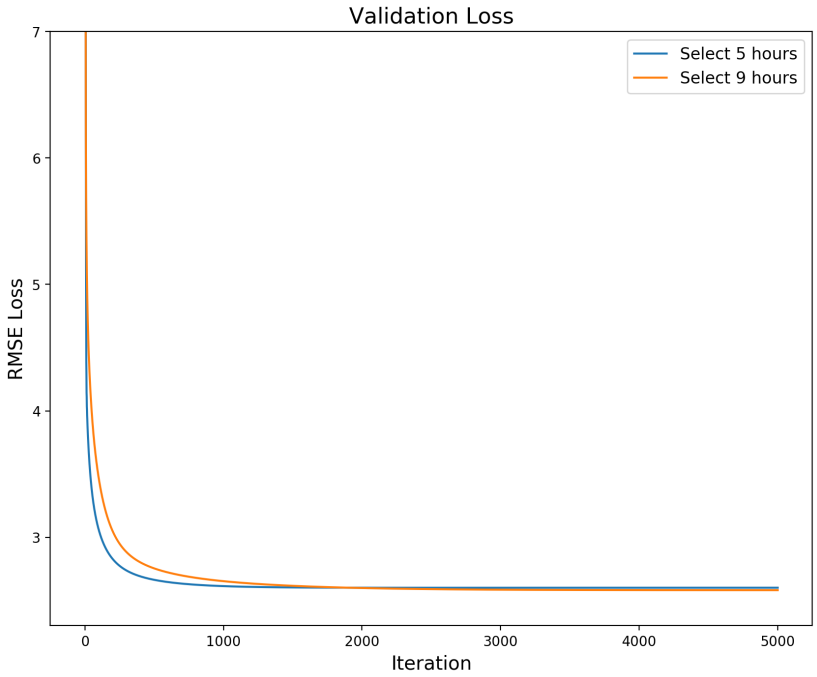
1. 使用四種不同的 learning rate 進行 training (其他參數需一致)，作圖並討論其收斂過程（橫軸為 iteration 次數，縱軸為 loss 的大小，四種 learning rate 的收斂線請以不同顏色呈現在一張圖裡做比較）。



從上圖可以發現 loss 的收斂速度由快到慢為 10, 1, 100, 0.1，由此可以發現收斂速度並不一定隨著 learning rate 增加而加快。而當 learning rate 設定過小（例如圖中的 0.1），雖然 loss 有持續遞減的跡象，但收斂速度比其他三個 learning rate 慢了非常多，且無法確定需要多少個 iteration 才能收斂。

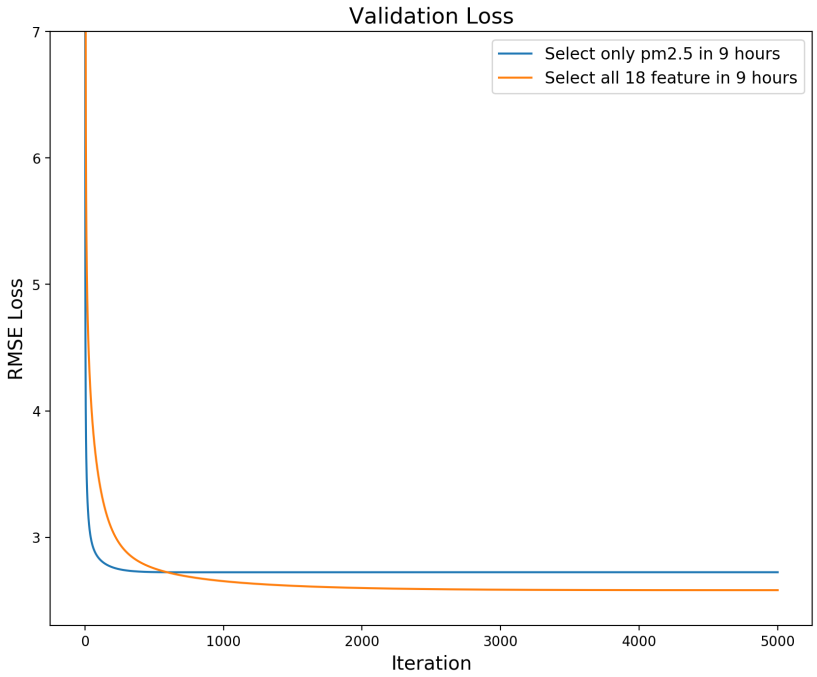
2. 比較取前 5 hrs 和前 9 hrs 的資料（5*18 + 1 v.s 9*18 + 1）在 validation set 上預測的結果，並說明造成的可能原因。

從下圖可以發現只取前 5 小時和取前 9 小時的資料在 validation set 上預測的結果非常近似，因此推測 5 小時內的觀測數值變動率足夠小，因此只要 5 小時內就可以預測出第 10 小時的 pm 2.5，或著是說前 9 個小時到前 6 個小時的資料對於預測第 10 個小時的 pm2.5 數值的影響非常小。



3. 比較只取前 9 hrs 的 PM2.5 和取所有前 9 hrs 的 features（9*1 + 1 vs. 9*18 + 1）在 validation set上預測的結果，並說明造成的可能原因。

   從下圖可以觀察到取所有前 9 小時的 features 的預測結果相對較好，但這個預測結果與只取前 9 小時的 pm2.5 的預測結果相差不大，由此可以推測雖然其他 17 個觀測值可以提供較多的資訊來預測，但大多數的 feature 可能與 pm2.5 數值的相關性不大，第十個小時的 pm2.5 數值主要還是與前 9 個小時的 pm2.5 數值有關。



4. 請說明你超越 baseline 的 model(最後選擇在Kaggle上提交的) 是如何實作的（例如：怎麼進行 feature selection, 有沒有做 pre-processing、learning rate 的調整、advanced gradient descent 技術、不同的 model 等等）。

   我的 model 仍然採用手刻 linear regression，在 feature selection 時選擇了所有 9 小時的 18 個觀測值。由於 pm2.5 的預測值與前 9 個小時的預測值有高度相關，而 pm2.5 中出現少數負數的異常值（pm2.5 應該要大於 0 才對），而考慮到觀測值的變動應為連續的，所以在 pre-processing 中將這些負數的 pm2.5 修改成前一個小時的數值（若前一個小時也是負數的，則繼續找前兩小時，以此類推），以此來減少異常的 pm2.5 數值對預測值的影響，最後再對所有的 18 個 feature 做 normalization。此外考慮到線性模型可能無法較好的解讀風向的資訊，因此將風向取 cosine 和 sine 值，由此代表風向的向量（x 與 y 方向）。

   在 model 的訓練上採用了 `Adagrad`，迭代次數 `iteration` 設定為 5000（確保有足夠的次數可以收斂），`learning rate` 設定為 50。接著實作 5-fold cross-validation，以避免出現 training 和 validation set 切割不均勻的狀況。在訓練完成時挑選 cross-validation 中 training 和 validation loss 差異最小的 model 作為 best model。